9 The practice of short-term climate prediction

1) seasonal mean, 2) lay-out, 3) time-scales, 4) elements/methods, 5) expressing uncertainty, 6) simplifications, 7)format of the forecast, 8) the official forecast, 9) verification I, 10) verification II, 11) role of trends, 12 frcst of opportunity. Appendix hist notes,

While previous chapters were about methods and their formal backgrounds, we here present a description of the process of making a forecast and the protocol surrounding it. A look in the kitchen. It is difficult to find literature on the subject, presumably because a real-time forecast is not a research project and potential authors (the forecasters) work in an ever-changing environment and may never feel the time is right to write an overview of what they are doing. Moreover, it may be very difficult to describe real time forecasts and present a complete picture. Nearly all of the material presented here specifically applies to the seasonal prediction made at the NWS in the USA, but should be relevant elsewhere.

A real time operational forecast setting lacks the logic and methodical approach one could strive for in science. This is for many reasons. There is pressure, time schedules are to be met, input data sets could be missing or incorrect, and one can feel the suspense, excitement and disappointment associated with a forecast in real time. There are habits that are carried over from years past - forecasters are partly set in their ways or find it difficult to make major changes in mid-stream. The interaction with the user influences the forecast, and/or the way the information is conveyed. Psychology enters the forecast. Assumptions about what users want or understand do play a role. Generally speaking a forecast is thus a mix of what is scientifically possible on the one hand and what is presumably useful to the customer on the other. The CPC/NWS forecasts are moreover for the general user, not one user specifically. Users for short-term climate forecasts range from the highly sophisticated (energy traders, selling of weather derivatives, hydrologists) via the (wo)man in the street to entertainment.

The seasonal forecast has been around a long time in the US. Jerome Namias started in-house seasonal forecasts at the NWS in 1958. After 15 years of testing, his successor Donald Gilman made the step to public release in 1973. The seasonal forecast had been preceded by a monthly forecast (starting in the early 1950's) which in turn was preceded by a 5 day mean forecast for days 2 through 6 (well before NWP played a role), a project that started around 1940 with several collaborators at MIT (Rossby, H. Willett, J. Namias). Some of the attributes of today's seasonal forecast, for instance the use of a three class system, date back to these early efforts around 1940 (Rossby 1941). For a few more historical notes, see appendix.

In this chapter we discuss a number of issues, specifically the rationale for time averaging, the lay-out and format of the forecast, the (in)famous three-class system, what is forecast and by which methods, a-priori and a-posteriori skill, hindcasts, the role of trends etc.

9.1 On the seasonal mean

Why are we forecasting the seasonal mean, an average over about 90 days? Users, when asked, may express a desire for daily forecasts out to infinity, but here the limits set by state of the art science prevail over user desires. It is impossible as of now to forecast, with skill, day by day weather beyond one or two weeks. Fig. 9.1 is an example of 500mb height verification out to 30 days in a 5 year forecast data set produced retroactively by a 2002 NCEP global NWP model (Jae Schemm, private communication). Beyond two weeks the correlation of daily forecasts with verifying analyses is small (<0.20), even in the leading modes NAO and PNA, but not completely zero either. Assuming the remaining correlation is worthwhile to some users a seasonal mean is taken as a filter to amplify the signal to noise (SN) ratio - note that many verification measures relate to the SN ratio (Compo and Sardeshmukh 2004). The signal is defined here as the predictable part of the weather, while the rest is noise. Forecasting the seasonal mean over day 15-105 ahead of time is thus somewhat of an admission up front that skill is inherently limited in that range. For similar reasons NWS's CPC has had a 6 to 10 day averaged forecast (since 1978) and a week2 forecast (since 1997) where low skill is addressed by taking a 5 and 7 day mean respectively. The lower the correlation (but still positive), the stronger the averaging one would need to reduce the noise sufficiently. But one also needs to be sure the signal does not get harmed by the averaging. (A time mean over the first week of NWP forecasts would be unwise because the signal (well forecast early on and time varying) is harmed by taking a time mean.) A time mean in a situation of nearly constant signal (i.e. constant with lead time) is, in purpose, comparable to taking the mean of a modern ensemble (Tracton and Kalnay 1993) - the purpose is to improve skill by some measure.

The transition from week2 to a season in terms of averaging length is rather abrupt. Indeed a monthly (mean) forecast in between week2 and seasonal may seem advisable. Currently, the intraseasonal forecast (say a monthly mean from day 15 to day 45) is still very difficult and has low skill, lower than the seasonal mean at longer lead. At ultra-long leads one could consider time averages longer than a seasonal mean, but here the user's needs prevail. Few users would be served by an annual mean forecast, even if it had some skill. (For similar reasons prediction of spatially averaged quantities are rarely considered practical (all weather is 'local'), even though in research 'all-India' rainfall has been the target of prediction (Mooley et al 1986)). The seasonal mean is probably the longest time average one can afford without mixing wildly different winter and summer climates.[1]

9.2 Lay-out of the forecasts

Fig. 9.2 shows a lay-out of the forecasts at NWS - time progresses towards the right. The day-by-day short-range forecasts (day 1-7, not discussed here) are followed by the 6-10day/week2 forecasts which are already statistical in nature in that the target is a time mean, and the format used is probabilistic (O'Lenic ref). Through week 2, i.e. through day 14, the basis of the forecast is almost entirely in NWP, with its suite of ensembles (Tracton and Kalnay 1993). The shortest seasonal forecast is applied to day 15-105 averaged - its 'lead time' is 2 weeks. (The lead time is defined as the amount of time between the moment a forecast is issued and the first

[1] OCN is the exception. While OCN is a ten or 15 year average, it is applied only to a certain target season.

moment of validity.) Then twelve more rolling seasons follow, each increasing the lead time by 1 month, out to 12.5 months, see Fig.9.2 for a schematic. For example in mid-November 2005 seasonal forecasts were prepared covering DJF2005/06 (lead 0.5 mo), JFM2006 (lead 1.5 mo)... through DJF2006/07 (lead 12.5 mo). This suite of forecasts is released every 3rd Thursday of the month. The DJF2005/06 forecast was first issued as a 12.5 mo lead forecast in November 2004.

At CPC the expression 'long-lead' is used. This refers to the fact that even the first seasonal forecast starts beyond week 2. In the past, prior to 1995, the seasonal forecast started at lead zero, i.e. started its validity immediately after release. The wisdom of having (or not) a zero lead seasonal and monthly forecast continues to be debated among users, forecasters and researchers.

9.3 Time scales in the seasonal forecast

While seasonal sounds like 90 days, the time scales that need to be kept in mind are several:

a) averaging time,

b) lead time and

c) time scale of physical processes that contribute to skill.

We again refer to the lay-out of the seasonal forecast in Fig.9.2 for explaining a) and b). The *averaging time* (our choice) is the easiest: 3 months. The *lead time* is the amount of time between the moment a forecast is issued and the first moment of validity. For instance a DJF forecast issued on November 15 has a 0.5 month lead. Lead time and averaging times are choices and could be altered if necessary. The *time scale c)* , beyond anybody's control, is related to *physical processes* in the geophysical system that can be tapped to make a forecast with some skill. Among the most important processes we have mentioned in previous chapters are ENSO (time scale many months to a few years), low frequency variations and trends (multi-year, decades or more time scale) and soil moisture effects (a few months). In principle a full spectrum of time scales is or could be relevant to the seasonal mean climate[2]. It may come as a surprise to some that multi-decadal low frequency variations (e.g. OCN, see Ch8.3) should play a big role in the seasonal forecast. This is a reflection of not only the strength of such trends, see sections below, but also the lack of skill at the shorter time scales of the spectrum.

9.4 Which elements are forecast, and by which methods?

Table 9.1 shows the elements being forecast officially at CPC and the methods used to accomplish this. The *elements* (left to right) consist of seasonal mean T&P, the sea-surface temperature and the continental soil moisture (w). For T&P the official forecasts are restricted to the US (including AK and HI), even if several tools are for the whole world. The SST is forecast for the whole world ocean, but the only official NWS SST forecast refers to Nino34. SST plays a role comparable to Z500 in short-range weather prediction, i.e. its skillful forecast is important because simultaneously occurring surface weather can be derived from it for the locale of interest. w is part of a pseudo official forecast only in that the Drought Outlook is largely based on and verified against w. Many other elements are forecast by tools, but have no official status.

Method\Element	US-T	US-P	SST	Soil Moisture (w)	References
CCA	X	X	X		Barnston(1994)
OCN	X	X			Huang et al(1996)
CFS	X	X	X		Saha et al(2006)
CA-SST			X		Van den Dool and Barnston(1994)
CA-w	X	X		X	Van den Dool et al(2003)
ENSO Composites	X	X			

[2] Time scales less than 90 days are also present in a 90 days mean (Madden 1976), and mainly to the detriment of skill.

Other Models	X	X	X		
Markov (MRK)			X		Xue et al (2000)
Consolidation	X	X	X	X	
MLR	X	X			Unger(1996)

Table 9.1 Overview of elements (left to right) and methods (top to bottom) that play a role in CPC's seasonal forecast.

Of the *methods* listed in Table 9.1 CCA, OCN, CA, ENSO, composites, CFS, MRK, MLR and Consolidation have all been discussed in Chapter 8. The CCA, OCN and CFS are the standard tools for US T&P used every month, while the CA for soil moisture and ENSO composites are examples of tools of opportunity, used only during the warm half of the year (soil moisture) and during ENSO winters (composites are invoked when forecasts for Nino34 indicate a warm or cold event). Any of the tools can be accessed in real time via http://www.cpc.ncep.noaa.gov/products/predictions/90day/tools/briefing/. The "other models" mentioned in Table 9.1 are imported from institutions outside NCEP, subject to timeliness, a-priori verification and other operational protocols etc. Including all members in each institution's ensemble the forecaster has access to order 100 different forecast, a formidable task for any human being. The method 'consolidation' was either (certainly in the past) a subjective process of combining all information, or (in the future) a largely objective combination of tools as described in Chapter 8.9.

Table 9.1 also shows some historical carry over and 'accidental' aspects. Instead of CCA for instance, a number of similar methods could have been developed, see chapter 8.7. This circumstance often depends on the preference and interest of personnel at a particular time. If a method listed in Table 9.1 is used to forecast only one or two elements, that does not imply it could not be used also to forecast the other elements, just that the research and development was not done.

CCA as a method was discussed in Chapter 8. The lay-out of the predictors of the original

CCA at CPC is such that global SST and 700mb height during the last 4 non-overlapping seasons are truncated by EOFs and compressed into a low dimensional predictor vector. The predictand is also heavily truncated before the CCA is done. The truncated predictand at an earlier time is part of the predictor. More recently, a variant called ensemble CCA (Mo 2003), ECCA, has been added, but for the first lead only. This ECCA is based on upper level velocity potential, soil moisture etc, but at one antecedent time level only. The MLR at CPC (Unger 1996) was developed to be a methodological alternative to CCA but with identically the same predictor-predictand lay-out. Soil moisture was added as an additional predictor.

In addition to the tools mentioned, there may be more informal aids. In fact the forecaster has a mental checklist that includes 'local' effects (especially SST anomalies along the south California coast), short-term persistence, the very latest on Nino34 (and an adjusted larger or smaller role for ENSO composites), opinions expressed in a monthly phone conference (often backed up by a researcher on the outside running a variation of an accepted tool) etc. In addition to individual tools, consolidated renditions of any 2 or 3 tools are available, for example ENSO composites combined with trend (Higgins et al 2004), CCA and OCN combined etc etc.

9.5 Expressing uncertainty

Because of its limited skill, it is important to express uncertainty for the seasonal forecast. Whatever little skill is available should not get lost in translation. This was recognized early (Gilman 1985), well before NWP had ensembles, and long before probability forecasts were an acceptable wide-spread practice in NWS (it still isn't!). It is apparently an article of faith that uncertainty shall be expressed through a probability forecast. One may think here of error bars (the standard deviation of a Gaussian distribution around the point value), or a complete probability density function (pdf). As shown in the example in Fig. 9.3, the forecast is basically thought of as a statement that nature will draw a realization from the conditional pdf (cpdf; red/dashed). (Note that Fig.9.3 features pdfs of seasonal mean values, not pdfs of daily values during the season). The word conditional refers to a pdf subject to the initial condition and all that is knowable about the future at that time in that circumstance. If there is no skill, and the forecaster understands this correctly, the forecast should match the climatological pdf (full black/blue line; as determined for instance from data over a standard climatology era, such as 1971-2000). In the example in Fig.9.3 the forecast has a warmer pdf (if this is temperature) by a noticeable shift to the right and suggestive of skill both by the shift in the mean and a narrower and higher distribution. As is also conveyed, a positive point forecast (the median of the distribution) in a situation with skill (a +0.5 shift and narrowed pdf) does not rule out a (high) negative value, just that the probability has been reduced greatly. More on Fig. 9.3 later.

{{We here present pdfs and probabilities as a means to express forecast uncertainty, but these concepts have a less than obvious intersection with the concepts weather and climate (prediction), see also the first footnote in Ch1. Diagnostically it makes perfect sense to define climate as the pdf, and weather as a single realization drawn from that pdf. When using cpdfs as a means to express forecast uncertainty it thus makes some (not perfect) sense to look upon a cpdf as a climate prediction, while deterministic single forecasts or any point forecast would be weather prediction. This nomenclature appears to be followed more or less these days. The logic is imperfect because the needs for pdfs occurs at much shorter leads for P than for T. Nobody has declared the probability for P for the next 12 hours a climate prediction.}}

While the shorter range forecasts may have largely escaped a formal probabilistic approach (temperatures in the low fifties), the longer range forecasts have excelled in following a precise probabilistic protocol. Main problem: will the public understand this? How to convey probability information (such as in Fig.9.3) understandably and correctly to a large audience is a subject of continuing discussion[3]. It seems obvious that many users would not appreciate a 'complete' pdf,

[3] Short of providing a complete pdf, uncertainty has been conveyed in several other ways. Maximum temperatures in the lower fifties has a purposeful uncertainty, i.e. the forecaster would not dare to say 52.4F. The use of explicit error bars is uncommon in meteorology, but the use of categories, and making categorical statements (temperature will be in the below normal class) is similar to 100% confidence bars (open ended on one side for the 2 outer classes).

whether it is provided as a graph, analytical function or a detailed tabulation for each locale. A simplification is needed, as described directly below.

9.6 Simplifications of the probability forecast (the three classes)

Instead of a complete pdf, which would in principle be an analytical expression plus the numerical values of the parameters describing it, the following simplifications have been used for over 20 years.

1) Three classes, or terciles, are used.[4] Based on the climatological pdf three classes can be defined, named Below Normal, Near Normal and Above Normal (B, N, and A) - in Fig.9.3 the vertical dashed lines at +/- 0.4308 (in standardized units) signify the tercile boundaries for a Gaussian distribution.[5]

2) By integration the conditional probabilities for B, N and A can be determined - in the example in Fig. 9.3, one finds 15%, 32% and 53% respectively. Indeed the probability for an above normal outcome has increased noticeably, mainly at the expense of the other extreme. The probability for the N class changes surprisingly little, unless the cpdf shifts are considerably larger than the half standard deviation used in the example.

At this point the simplified probability forecast, at each locale, consists of three numbers (p_B p_N and p_A) which, since they add up to 100 could be given by two numbers per location, still a complex map, i.e., 2 maps collapsed into one. The desire to present information as a simple understandable national map for public consumption (like any other weather map) forces another simplification:

3) $p_N = E$ and $(p_B - E) = -(p_A - E)$, where E is the climatological probability (1/3rd).

[4] The three class system for categorical forecasts is at least 65 years old (Rossby 1941). Three class probabilities were introduced in 1982, see Gilman(1985).

[5] The discussion is easiest for a Gaussian distribution, but three classes can be defined for nearly any distribution. CPC uses a 2 parameter gamma distribution for P.

Equivalently $p'_B = - p'_A$ and $p'_N = 0$, where p' is the probability anomaly.

The seasonal forecast maps issued by CPC show contours of p'_B or p'_A, whichever is positive, with E added back in[6]. In the absence of any skillful information about the future the forecast, labeled nowadays EC (equal chances) would be 1/3rd, 1/3rd, 1/3rd.

For the advanced user the (more) complete pdf can be accessed in tabular or graphical form for many locations in the US - and none of the above simplification is used. (See http://www.cpc.ncep.noaa.gov/pacdir/NFORdir/HOME3.shtml for real time examples)

The use of three classes, while meant as a simplification, also creates an array of problems and questions.

1) Why three classes, and why three *equal* classes? Clearly, a large number of classes is in the limit the same as a complete pdf, so, in order to simplify we need to reduce to just a few classes. Low skill also argues in favor of only very few classes. An odd number of classes would seem preferable as it leaves the neutral middle, the maximum of the pdf, as one entity. Some organizations have however used two classes, cutting the pdf in two parts right at the median, thus forecasting only the sign of the anomaly.[7] Three classes is thus the lowest number of classes that, in our opinon, makes sense as a simplification of the full cpdf. Until 1995 CPC used to have three classes based on a 30/40/30 climatological distribution. The wider N class was implemented to combat the lack of skill in the N class, a, by now, well understood problem (Van den Dool and Toth 1991), but the unequal classes always raised questions with users. In 1995 we went to three 'equal' classes. Here is another reason why equal classes simplify: the notion 'equal chances' would make no sense if the (30,40,30) classes are used. Indeed, in the past we have used various other symbols for EC: I (indeterminate), and CP and CL (both meaning climatological probabilities).

[6] For nearly 10 years we made maps of probability anomalies, but went back to full probabilities recently at user request.

[7] At times CPC has had a two class version of the three class forecast

2)The three classes have become so much the public face of the forecast that many people, even insiders, appear to have forgotten that it is meant as a simplification.

3) Another mystery to many is ‘the event’. Probability forecasts tend to be, in statistical parlance, for ‘an event’. When the event is rainfall (or being hit by lightning) most people understand the concept, because it rains or it does not rain. The 50-50 concept as it relates to the flip of a coin is also widely understood. However, when the event is temperature falling in one of three terciles, the abstraction level is suddenly a challenge and ‘the event’ somewhat mysterious. Explaining the situation with dice, or, by abstraction, a three sided die might help. One could say that if nature throws the loaded three sided die in the example in Fig. 9.3 an infinity of times, the B, N and A sides would appear 15, 32 and 58% of the time[8]. Clarification by invoking concepts in gambling (the odds[9]), flip of a coin, while highly applicable, is not uniformly appreciated by management at all times, because it suggests a non-serious activity.

4) With modest probability shifts it frequently happens that the most favored class has less than 50% chance being categorically correct. By implication the favored class is more likely wrong than right. This causes bewilderment. (For these users the 2-class system may be better).

5) There are in general negative connotations associated with any probability forecast. To many it seems as though we are seeking a formulation to never be completely wrong. By the same token a probability forecast is never a complete hit, unless one places 100% probability in the correct bin.

9.7 Format of the forecast

Fig. 9.4 shows an example of a set of forecasts released to the public in the middle of March 2006. These are the 0.5 month lead seasonal forecasts for AMJ 2006. Temperature is on the left, Precipitation on the right. There are basically four options the forecasters have at their

[8] Never mind that nature does it only once

[9] For many years we were not allowed to use the word odds.

disposal to fill in these maps:

a) A shift of probabilities towards above median, as in Fig.9.3, and as shown in Fig. 9.4 in much of the southwestern US for AMJ 2006 for T. The contours, 33, 40, 50, etc (with 1/3rd subtracted) indicate how much the probability shift to the above median tercile amounted to. A suggestive color is used: orange-red (green) for above median[10] T (P).

b) The same as a) but now a shift of probabilities towards below median. Here the colors are blue for T (as in northwestern US) and brown for P (southern states).

c) Equal Chances (EC); $p_B = p_N = p_A$ = 1/3rd. This would be areas left blank where no single tool has non-zero a-priori skill, or signals by various tools with alleged skill are in conflict. EC is an informed "we don't know".

d) An option (not used in Fig. 9.4) for enhanced probabilities of the Near Normal class. Occasionally we give an N5, meaning that we borrow 2.5% from both extremes to make the distribution higher and narrower, but no shift. This would happen in an area with very high skill (in general) but a low signal on a particular occasion[11], and also if two high skill tools give opposite forecasts. The N option is rarely used because skill is so low for Near Normal (Van den Dool and Toth 1991; Kharin and Zwiers 2002). This is caused mainly by the unfavorable ratio of the width of the class to the rmse (the error bars) - in a dry climate one has the same problem with the B class on P being very narrow (and one shower kills a forecast for B).

Under options a), b) and d) only positive contours are shown. The reader is supposed to know the implied negative probability anomalies for the other classes, as described in here.

We have to accept this reality: Many users, and even some insiders, will simply look at the color, forget the contours, the pdf and the assumptions, and convert the map into categorical

[10] We use the notion median instead of mean because the precipitation is skewed. Median and mean is the same, or very nearly so, for seasonal T.

[11] For instance, if Nino34 correlates very highly with seasonal temperature at a locale of interest, then the chances of either extreme class to occur are reduced in a neutral year.

forecasts. Orange is thus above normal, green is above median etc etc, and forecasts will be judged categorically.

The colored areas on the maps are sometimes referred to as non-EC.

The opposite of option d, a wider pdf with reduced (enhanced) probability for N (outer classes) is technically possible but not practiced in the official forecast.

9.8 The official forecast

Table 9.1 is just a listing of 'tools' to make the official forecast. But how is the official made? It is convenient to think of the official forecast as a (linear) combination of the tools, e.g.

$$OFF= aA+bB...zZ \ , \qquad (9.1)$$

where the capitals refer to methods (CCA, CFS etc), and the lower case coefficients depend on skill of each method and the co-linearity among them, see Chapter 8.9. Assuming we know the skill of the forecast (from many hindcasts, see next section) we need to convert Eq (9.1), a point forecast, to a probability forecast. For the three class system this can be done directly, in simplest form, as per figures like Fig. 9.5, which show the probability anomalies (p') for the extreme class (say the A class) as a function of a) a-priori skill (expressed as a correlation labeled R), and b) the departure of the point forecast from climatology (=shift of the cpdf, labelled F). Fig. 9.5 was prepared by David Unger. As expected, probability anomalies increase with both the correlation (R, in the vertical) and the strength of the point forecast (F, in the horizontal). While this is qualitatively quite obvious, Fig. 9.5 provides a quantitative conversion depending on two knowable factors. These two factors, R and F, are not totally independent of course. In a situation of zero R the anomaly point forecast should have been damped to zero. But for a modest non-zero R of say 0.5, the value of F, when extreme, can make a large difference in the probability. The same graph can be used for both A and B (but note the asymmetry relative to F=0), and the remainder for N then follows. For large R and F the p' for the extreme class is more than E

(1/3rd) - at this point one of the simplifications ($p'_A = - p'_B$) we described in sct(9.6) can no longer be applied and one would need to rob points from the N class as well as from the opposite extreme.

Because the suite of 13 seasonal forecasts is made each month, a certain target season at lead τ (except the very last τ=12.5months) already has last month's official forecast at lead τ+1 as first guess - this way corporate expertise is handed down for 12 months in a row until the final opinion at the shortest lead τ=0.5 is issued. So (9.1) could be written:

OFF= first guess + aA+bB...zZ + subjectivity, (9.1a)

The subjectivity should be kept to a minimum.

9.9 Verification I - a-priori skill (Hindcasts)

Verification has been mentioned already several times, and indeed, verification is part and parcel of any credible forecast operation. In short-term climate prediction two kinds of verification exist. The obvious is the a-posteriori verification - after the fact one wants to determine the skill of (a particular set of) forecasts (see sct (9.10)). Less obvious is the so-called a-priori verification. The latter was developed to address situations with modest skill and/or situations where forecasts are issued infrequently. In these cases it is of paramount importance to give the user a sense of how much faith to put in the forecast. A forecast without some sense of a-priori skill can do more harm than good. In the short range weather forecast there are very many forecasts in quick succession in real time that may give the user an impression of the skill level, which can then be mentally applied to the next forecast by the user, but in short-term climate prediction there are only a few independent forecasts per year. No-one remembers far enough back to accumulate sufficient statistics of real time forecasts for, say, DJF. One alternative is to evaluate a (large enough) set of retroactive forecasts (also called hindcasts) which mimic the real time situation as faithfully as possible. Around 1990 this approach was first in place for several individual statistical objective tools that aided in the seasonal forecast. The need for hindcasts has increased exponentially since then because the number of tools is increasing very quickly. The relative weights of forecast tools in Eq(9.1) cannot be determined unless there are (enough) hindcasts to base them on. This is especially so when co-linearity among tools is large, see discussion on Consolidation in Ch8.9. Hindcasts are also needed to bias correct, calibrate (probabilities) and verify each tool in its own right. Hindcasts can only be made for objective tools. Subjective forecasts and even the official forecasts cannot be credibly rerun over the last 25 or 50 years.

For statistical-empirical tools developing a set of hindcasts is easy, in principle, and can be done for a period covering the length of the data sets involved (~ 50 years). To make multi-membered hindcasts for a dynamical coupled ocean-atmosphere model the investment is much larger, and demands on CPU very high. Moreover, reliable initial conditions for the global land and ocean as required by dynamical models may not (yet) exist or may always be impossible given data scarcity, especially before 1980 (ocean). Because a consolidation is most easily based on the common period of the hindcasts, the operations at CPC and NCEP uses the period 1981-present for the hindcasts (even if normals are 1971-2000). The need for hindcasts has increased the need for observations, the need for recovery of nearly forgotten observations and the need for state of the art global reanalyses of which Kalnay et al(1996) and Kistler et al(2001) was only the beginning.

Hindcasts, if affordable, have this major advantage. Every time a tool is changed a new set of hindcasts would be available and there is no need to wait months or years before making an assessment of the skill of the new tools in real time operations. This assumes a set of diagnostics and verification can be run instantly.

Even for statistical tools a hindcast data set cannot be obtained without some investment. Statistical tools may suffer from overfitting and give a too optimistic view of skill to be expected in subsequent real time forecasts. In view of a general impatience among clientele and funding agents, the old way of making forecast in real time and waiting until one has a sufficiently sized data set for evaluation will not do any more. Hindcasts are thus the approach of choice. To combat overfitting in hindcasts 'cross-validation' has been invented. In that procedure one or (better) several years are left out, and we act as though they never occurred. The statistical model is developed on the retained years, then applied to the withheld year(s). This is done exhaustively for each year withheld. In some strategies one needs more than one level (i.e. nested) of cross-validation. In this way even a statistical tool may consume a lot of CPU. The science of cross-validation itself has to be further developed - several downsides and boobytraps have been noted (Barnston and Van den Dool 1993).

A forecast like the one in Fig 9.3 is thus based on both a real time aspects (F, the strength of the predictor) and hindcast aspects (R the a-priori correlation). Fig.9.6 shows a rendition of the real time forecast by the CCA tool for the entire US for MAM 2006. (Other tools are presented in the same way to the forecaster). Here, in a nutshell, both the current state, and the hindcasts over 1955-last year play a role. The point forecast (in units of standard deviationX10) is strongly non-zero if a) CCA had, locally, skill over the past 50 years and b) the predictor (modes of global SST mainly) are sufficiently anomalous. The cutoff for local skill is taken to be 0.3 - any correlation below 0.3 is considered indistinguishable from zero, practically or statistically.

9.10 Verification II - Heidke Skill Scores.

There is no substitute for making forecasts in real time and doing verification a-posteriori. Here we report on a verification of the official (OFF) seasonal temperature forecast over 1995-2002, all 102 Climate Divisions, all 13 leads and all 12 initial seasons combined. The measure used, mainly for the sake of tradition, is the Heidke Skill score, a categorical score generically defined as:

$$SS = (H - E) / (T - E) * 100 \qquad (9.2),$$

where T = total # of forecasts, E=T/3 is the expected number of hits by chance, H = # of categorical hits. Because we make EC forecasts it makes sense to verify the non-EC forecasts (the colored areas in Fig 9.4) first as:

$$SS_1 = (H_1 - E_1) / (T_1 - E_1) * 100 \quad (9.2a),$$

where T_1 is the total # of non-EC forecasts, $E_1 = T_1/3$, and H_1 the number correct. Coverage is defined as T_1 / T. A score for the whole nation, including the blanks (the EC area), is produced by counting EC forecast as 1/3rd correct. One simply finds: $SS_2 = SS_1$ * coverage.

How to judge a Heidke score of 20-25%? For readers more familiar with correlations: On a large set of forecasts, and modest skill, the Heidke score for a three class system equals half the correlation (Barnston 1991) - e.g. SS=20 corresponds to a 40% correlation. Relative to a probability forecast, one can convert by considering that on average the observed class has been forecast as the favored class 13-17% more often than expected by chance. I.e. the average probability shift is comparable to what is shown in the example in Fig 9.3.

Table 9.2 Heidke Skill Score of CPC Temperature Seasonal Forecasts for JFM95-FMA2002. All 102 climate divisions, starting times and lead are combined. The CCA and OCN methods were unchanged during this period, while the dynamical method (predecessor of current CFS) changed several times.

	SS_1	SS_2	Coverage	
OFF	22.7	9.4	41.4%	(13 leads)
CCA	25.1	6.4	25.5	
OCN	22.2	8.3	37.4	
Dynamical	7.6	2.5	32.7 (1st 4 leads only)	

In Table 9.2 we only show tools that are used all the time, and have been archived from the beginning of the long lead prediction in late 1994. In order to compare the performance of two tools one is advised to compare SS_2 . We thus conclude from Table 2 that the official forecast is better than the participating tools. This is mainly from increased coverage. Apparently CCA and OCN, while having similar SS_1, have skill in non-overlapping areas and the forecaster is capable of combining these two tools to arrive at a superior official forecast. (Keep in mind that the forecaster, on occasion, uses other tools, such as ENSO composites which were highly successful in winter 1997/98, see Barnston et al 1999). SS_1 is important to verify that the a-priori skill estimates used in real time (but based on the historical record up to that time) were correct. Considering that we use a 0.3 correlation as cut-off one wants SS_1 to be in excess of 15. CCA and OCN's a-priori skill estimates appear to be correct and holding up on independent data. The dynamical model used over 1995-2002 may not have been optimal. The a-priori skill estimate was inaccurate, or the model in real time was not exactly the same as the one used for the hindcasts. We believe this has improved since summer 2004 when the CFS was introduced (Saha et al 2006). On the whole CPC makes non-EC forecasts for about 40% of the nation. While this may seem a downer for the remaining 60%, one needs to keep in mind that it boosts faith in the non-EC forecast where and when they are issued. Forcing forecasters to make a non-EC forecast under all circumstances, especially when skill is certified low, is counterproductive.

A table like Table 9.2 for precipitation is not shown because all SS values are very low, between 0 and 5, dangerously close to a random forecast. An analysis as to which tool contributes the most seems meaningless. If it were not for an occasional strong ENSO winter the skill of precipitation might indeed be very close to zero. An analysis of OCN over the years 1962-present shows that OCN-skill for precipitation in the 1960's through 1980's was generally better than it has been since 1995. Conversely, OCN skill for temperature was dangerously low in the 1980's when a regime of generally cold temperatures (all seasons in 60s and 70s) was replaced by generally warmer temperatures after 1995, first in winter and to lesser extent in summer.

Much more detailed regional verification is forthcoming, see Halpert and Pelman(2004).

What Table 9.2 does not convey is that nearly all skill in temperature is due to shifting probabilities to above normal for temperature. This point is addressed in the next section.

9.11 Trends

We have mentioned low frequency variations as a tool in making seasonal forecasts, especially for temperature, see OCN in Ch8.3. Table 9.2 shows that OCN has a strong contribution to the skill of the official forecast. The presence of trends has also a strong influence on the operational forecast in several other ways. Consider these facts: averaged across the United States temperatures for the 102 Climate Divisions over 1991-2005 have averaged 1.3 1.2 .6 .2 .4 .1 .2 .4 .5 .2 .1 1.0 C above the 1961-1990 mean for the months January, February, March... through December respectively. When expressed as (local) standardized monthly data before taking the national mean these shifts are .5 .5 .3 .2 .4 .1 .2 .4 .4 .1 .0 .4 respectively for the 12 months.[12] Had we known this in advance (in was not!), probability shifts on the order of what is shown in Fig 9.3, which was just an educational example, could have been expected for virtual all seasons just based on 'trends'. (The skill of OCN suggests persistence of 10 year averaged anomalies as a workable tool in real time.) In some areas, like the SW US the shifts are stronger, while they are weaker in the northern plains, see Fig. 8.2. One nowadays needs very strong interannual indications for a cold outcome in order to even dare to favor B. Keep in mind that many users round off the probability forecasts to a categorical (the one suggested by the color). Given how infrequent the B class is observed, for instance 7% instead of 33.3% of the time in 2005, see Table 9.3, forecasters nowadays shy away from placing high odds in the B class.

Table 9.3: The observed frequency (%) of occurrence of the three terciles in seasonal mean temperatures across the US.

B	N	A	at 102 US locations
(assumed to be 1/3rd, 1/3rd, 1/3rd, based on 30 year 1961-1990 normals period)			
26	28	46	1995
36	34	30	1996

[12] Introducing new normals in May 2001 has hardly lessened the bias. The difference of the 2001-2005 average relative to updated normals 1971-2000 is 1.0 -.1 .0 .7 .2 .2 .6 .4 .6 .4 1.2 1.0 C for the 12 months.

27	32	41	1997	
08	17	75	1998	
13	24	63	1999	
22	20	58	2000	
15	32	53	2001	(Normals changed! To 1971-2000)
19	36	46	2002	
15	38	47	2003	
20	33	47	2004	
07	32	61	2005	

During 1995-1997 the observed frequency was not very different from expected. But from 1998 onward (and in spite of the earliest possible update to 1971-2000 normals in May 2001) the outcome has been predominantly A.

Of the 4 options, favoring either B, N, A, or EC (climatological probabilities), only A and EC are used frequently, thereby calling the three class system into question. And forecast maps tend to look alike, regardless of lead, and to a lesser extent, regardless of season. It is only at the subtler level of probabilities that one can see the interannual component (due to ENSO or soil moisture) reduce the odds for above normal temperature that would be suggested by the trend alone. Managing this situation is a challenge. And an unannounced occasional cold month (with noteworthy societal impact) comes across as a huge bust. Given that in the 1960's and 70's the trend was for persistent cold (relative to 30 year normals in effect at that time), see Gilman(1986), the forecasters today wonder when the current warming trend is going to turn around. Often, so far erroneously, they feel it could be ‘now’.

9.12 Forecasts of opportunity and the tension with regularly scheduled operations

The level of skill reported in this book is not terribly high. Moreover, practioners know that a modest overall 0.35-0.50 correlation often hides a simple truth, namely that on a few occasions we have some truly usable forecast skill and in the rest of the circumstances we have virtually no skill at all. Following this point of view into the extreme we should perhaps refrain from issuing forecasts, except when the opportunity looks good, for instance when there is a strong ENSO event coming. The idea of 'forecasts of opportunity' is certainly not new, but is somewhat at odds with a regularly scheduled official forecast. Once a new forecast is expected each month it is hard to say: No, not now. The audience may no longer be there, when 5 years from now we finally see an opportunity. The way to manage this is by probabilities and in particular by using EC without being ashamed, and to make non-EC forecasts when and where the opportunity exists. In practice, however, there is considerable pressure to make non-EC forecasts more often, if not all of the time. For instance NOAA organizes press releases and conferences in the spring and fall as part of its annual activities calendar. A coast-to-coast EC forecast may not strike the audience as a great contribution to a newsworthy event, so the temptation is to put something on the map. This practicality has to be balanced against a more academic stand about forecast skill, credibility, and when and where we go for high probabilities.

Appendix: Historical notes

If someone wanted to describe how short-term climate prediction in the US is done in practice over the years, the literature would be helpful but quite limited. Chapter 9 describes the nuts and bolts of the seasonal forecast at CPC over the last 5 or 10 years, while chapter 8 includes the methods used at CPC formally during this period. One may have to go back to Wagner (1989) for a previous review with some detail. Some of the motivation about going to long-lead forecasts in 1994 is given in a trio of workshop papers (Barnston 1994; Van den Dool 1994 and O'Lenic (1994). From Gilman(1985; 1986) and Epstein(1988) one may surmise how the forecast in the US was made about 20 years ago. Gilmans's predecessor, Namias, was a prolific writer, and the period of the 1950's, monthly forecasts mainly, has been described quite extensively (Namias 1953). (Via Roads(1986) one can access more history about the Namias era, including Namias' collected works.) Van den Dool and Gilman (2004) summarized the influence of 50 years of NWP on the monthly/seasonal forecast. Finally there is a booklet for the 25th anniversary of CPC (Reeves and Gemmill 2004) with personal accounts by many of the forecasters and an attempt to write formal history (Reeves et al 2005).